AWS ML Associate — Section 1.2: Transform and Prepare Data

This section focuses on cleaning, transforming, and preparing datasets so they’re ready for ML workflows in SageMaker. It covers data wrangling, feature engineering, dataset splitting, and ensuring data quality.

# Data Transformation & Preparation Tools

• AWS Glue DataBrew — no-code visual data prep tool for profiling, transformations, and cleaning

• AWS Glue ETL Jobs — Python/Scala ETL scripts for large-scale transformations

• SageMaker Processing — run preprocessing code (Pandas, Spark, SKLearn) inside managed containers

• SageMaker Data Wrangler — visual tool for exploration and feature engineering, integrates with pipelines

# Data Preparation Concepts (Exam Focus)

• Data Cleaning — handle missing values (drop, impute), remove duplicates, normalize/standardize values

• Feature Engineering — encoding (one-hot, embeddings), scaling (min-max, z-score), text feature extraction (TF-IDF, embeddings)

• Splitting Data — train/validation/test (70/15/15 or 80/20), stratified sampling for classification

• Data Quality — detect imbalance or skew, apply oversampling/undersampling/SMOTE

# Integration with SageMaker

• Processing Jobs — scalable preprocessing before training

• Pipelines — automate transform → train → evaluate → deploy

• Feature Store — central repository for consistent features across training/inference

# Exam Tips

• Data Wrangler = preferred tool for feature prep in SageMaker

• Glue DataBrew = no-code prep; Glue ETL = large-scale code-based prep

• SageMaker Processing Jobs = flexible, scalable preprocessing environment

• Always split data properly to prevent leakage (train/val/test)

• Normalize/standardize numerical values for scale-sensitive algorithms

• Stratified sampling helps with class imbalance

• Use Feature Store for consistency across training and serving